Validity Arguments for EPA-Based Assessment

Validity Evidence for Assessing Entrustable Professional Activities During Undergraduate Medical Education
Claudio Violato, PhD, Michael J. Cullen, PhD, Robert Englander, MD, MPH, Katherine E. Murray, MD, Patricia M. Hobday, MD, Emily Borman-Shoap, MD, and Ozge Ersan, PhD

Abstract

Purpose
To explore validity evidence for the use of entrustable professional activities (EPAs) as an assessment framework in medical education.

Method
Formative assessments on the 13 Core EPAs for entering residency were collected for 4 cohorts of students over a 9- to 12-month longitudinal integrated clerkship as part of the Education in Pediatrics Across the Continuum pilot at the University of Minnesota Medical School. The students requested assessments from clinical supervisors based on direct observation while engaging in patient care together. Based on each observation, the faculty member rated the student on a 9-point scale corresponding to levels of supervision required. Six EPAs were included in the present analyses. Student ratings were depicted as curves describing their performance over time; regression models were employed to fit the curves. The unit of analyses for the learning curves was observations rather than individual students.

Results
(1) Frequent assessments on EPAs provided a developmental picture of competence consistent with the negative exponential learning curve theory; (2) This finding was true across a variety of EPAs and across students; and (3) The time to attain the threshold level of performance on the EPA for entrustment varied by student and EPA.

Conclusions
The results provide validity evidence for an EPA-based program of assessment. Students assessed using multiple observations performing the Core EPAs for entering residency demonstrate classic developmental progression toward the desired level of competence resulting in entrustment decisions. Future work with larger data samples will allow further psychometric analyses of assessment of EPAs.

Competence is a complex, interrelated, multidimensional construct to be acquired by medical professionals for safe and effective clinical practice.1–2 The complexities in the definition and structure of clinical competence, such as the excessive granularity of its measurable aspects, make competency assessment a challenge. One potential solution to this challenge is the use of entrustable professional activities (EPAs), which provide a holistic approach to competency assessment.3–4 EPAs provide a framework for assessment in competency-based medical education by requiring the integration of multiple clinical competencies in the authentic clinical environment.5–6 EPAs also provide a more concrete framework for assessors than the abstract, context-independent notion of a competency. In the aggregate, EPAs can provide predictability for future doctors’ practices by ensuring they are at the level of clinical competence expected when they start their residency or practice.6–10 They thus provide an assessment framework with content validity evidence.11–14

In 2014, the Association of American Medical Colleges published the Core EPAs for entering residency4 (Core EPAs). These Core EPAs were developed as a potential guideline for curriculum and assessment in undergraduate medical education (UME). They represent the essential tasks that all medical students should be able to perform on entrance to residency, regardless of specialty choice, without direct supervision. Further validity evidence will be critical to the widespread use of the Core EPAs.

Knowing how competence develops is critical to understanding whether an assessment framework for that competence possesses validity evidence. Validity evidence—particularly construct validity—for an assessment system such as EPAs can be derived from the demonstration that growth of performance with experience follows a theoretically predictable pattern.15 Empirical evidence that conforms to the predicted trajectory of learning based on theory provides evidence of construct validity of that system. Thus, we begin with a brief understanding of learning curves.16–18

Over 100 years ago, Thurstone posited that the relationship between the duration of practice and learning would be defined by a negative exponential function.18 If true, he predicted that gains in learning would be steep initially, but then plateau after some time as knowledge and skill acquisition takes place. Thurstone allowed that learning rates may vary depending on the type of skill or knowledge learned. Learning curves—graphical representations of the relationship between performance and experience—that demonstrate the relationship between deliberate effort...
and learning outcomes have been only sporadically employed in medical education. They illustrate the rate at which domain knowledge and skills are learned and also how quickly learning levels off.

Within medical education, learning curves have application for both individuals and groups. At the individual level, they illustrate the developmental trajectory for each student for different sets of clinical skills. At the group level, they illustrate the length of time it takes learners, on average, to attain a threshold level of knowledge or performance in executing a complex medical skill. Thus, learning curves have a variety of potential uses in medical education, including informing decisions regarding: (1) the duration of a program required for learners to acquire a predefined competency level; (2) how to use educational resources efficiently to reach the individual and program goals; and (3) how to enhance the curricula, teaching, assessments, and feedback to the learners.

The main purpose of the present study was to report preliminary validity evidence in using the Core EPAs as a framework for assessment. Data that demonstrate learning curves for both individual learners and aggregates of learners consistent with a negative exponential growth of learning provide validity evidence for the EPA framework. Accordingly, we analyzed the fit of learning curve models of negative exponential growth to repeated measures of performance based on EPAs. In a companion paper, we describe the operationalization of the program of assessment based on the EPAs at the University of Minnesota.

Method

Setting and participants

The University of Minnesota is part of a 4-site pilot, Education in Pediatrics Across the Continuum (EPAC), designed to test the feasibility of advancement from UME to residency and then from residency to fellowship or practice based on the demonstration of competence in a time-variable manner. Competency in this pilot was defined by an EPA framework of assessment, using the Core EPAs for the UME-to-residency transition and the General Pediatrics EPAs for the residency-to-fellowship or practice transition. This was a multiyear study. The data come from assessments generated by 14 students in 4 cohorts, who participated in the EPAC program during academic year (AY) 2014–2015 through AY 2018–2019 (the numbers of students from the 4 cohorts were 4, 3, 3, and 4, respectively).

Instrument and data collection

The EPAC longitudinal integrated clerkship program takes place during the third year of medical school. Students begin in June and continue until they have reached entrustment on the 13 Core EPAs, where entrustment is defined as the ability to perform the Core EPA with indirect supervision, with the supervisor checking all findings. At that time, they are able to progress to a transition phase of their education before the transition to residency. For monitoring the students’ growth on the EPAs, performance assessments were conducted on the 13 Core EPAs described in Table 1. Assessors (faculty and residents) rated the students on a scale from 1 to 9 (see Table 2) adapted from a supervision scale published by Chen et al.

Data analysis

The unit of analysis for this study is the number of assessments completed at the level of the EPA. Strictly speaking, the entrustment rating scale is ordinal, similar to the widely used Likert scale. An analytic approach employed for these scales is based on the assumption that they are interval or “quasi-interval.” This allows for robust parametric analyses such as Pearson correlation and multiple regression. Numerous Monte Carlo studies support the use of these parametric approaches in the field of ordinal analysis. The resulting estimate of the population parameter is almost always the best in the sense of being closest to the true value. Accordingly, we employed parametric analyses with the entrustment rating scale. Data were plotted over time, and curves were fitted theoretically based on regression coefficients derived from hierarchical regression analyses (Appendix 1).

A power analysis for the minimum required sample size for the multiple regression analyses described in Appendix 1 at the statistical power level <.05 required a minimum sample size equal to 84 observations. Due to the scarcity of data for some of the EPAs (i.e., < 85), we included only 6 of them in our analyses, all of which had 85 or more observations.

This study was determined to be exempt from review by the University of Minnesota–Twin Cities Institutional Review Board.

Results

Across the 4 years of this study, more than 2,400 total EPA-based assessments were made for the 14 students. Visual displays of 6 EPAs’ growth over the

| Table 1 |

<table>
<thead>
<tr>
<th>Description of 13 Core Entrustable Professional Activities (EPAs)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>EPA</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPA 1</td>
<td>Gather a history and perform a physical examination</td>
</tr>
<tr>
<td>EPA 2</td>
<td>Prioritize a differential diagnosis following a clinical encounter</td>
</tr>
<tr>
<td>EPA 3</td>
<td>Recommend and interpret common diagnostic and screening tests</td>
</tr>
<tr>
<td>EPA 4</td>
<td>Enter and document appropriate prescriptions</td>
</tr>
<tr>
<td>EPA 5</td>
<td>Document a clinical encounter in the patient record</td>
</tr>
<tr>
<td>EPA 6</td>
<td>Provide an oral presentation of a clinical encounter</td>
</tr>
<tr>
<td>EPA 7</td>
<td>Form clinical questions and retrieve evidence to advance patient care</td>
</tr>
<tr>
<td>EPA 8</td>
<td>Give or receive a patient handover to transition care responsibility</td>
</tr>
<tr>
<td>EPA 9</td>
<td>Collaborate as a member of an interprofessional team</td>
</tr>
<tr>
<td>EPA 10</td>
<td>Recognize a patient requiring urgent or emergent care and initiate evaluation</td>
</tr>
<tr>
<td>EPA 11</td>
<td>Obtain informed consent for tests and/or procedures</td>
</tr>
<tr>
<td>EPA 12</td>
<td>Perform general procedures of a physician</td>
</tr>
<tr>
<td>EPA 13</td>
<td>Identify system failures and contribute to a culture of safety and improvement</td>
</tr>
</tbody>
</table>
Validity Arguments for EPA-Based Assessment

10 months are presented as learning curves in Figures 1 and 2. These figures demonstrate rapid learning as students exceed a rating of 5 on the supervision scale (the cutoff score for entrustment to the transition phase to residency) corresponding to a 3a level in Table 2 defined as: "Is allowed to practice EPA only under reactive/on-demand supervision available and key findings double checked." In many instances, the students rapidly even reached a rating of 6 defined as: "Is allowed to practice EPA only under reactive/on demand supervision available and key findings double checked."

Ratings for the students across EPAs vary between 2 ("not allowed to practice EPA—allowed to observe") and 9 ("allowed to supervise others in practice of EPA"). There were only 2 instances of 9—one each for EPAs 1 (Gather a history and perform a physical examination) and 6 (Provide an oral presentation of a clinical encounter).

Figure 1 shows the trends of EPA raw data ratings for individual students over months. For any individual student represented by a line in the graph in Figure 1, there may be breaks in the line. These breaks correspond to no ratings on that EPA for the student represented by the curve in that particular month. In addition, at some time points, ratings across students are highly variable. Accordingly, the variance across the time points was not homogeneous (i.e., heteroscedasticity). Similarly, the standard errors might be different across the months. Therefore, the curves were "smoothed" based on regression modeling (Appendix 1).

Figure 2 shows the growth trajectories for all students combined for 6 EPAs over time, from June of the year they start to April the following year. These theoretically fitted curves suggest curvilinear relationships between months and EPA ratings, and hierarchical regression analyses suggest that for all EPAs, a curvilinear model was a better fit to the data than a linear model (see Appendix 1 for a description of these analyses). Figure 2 contains all the curves, along with 95% confidence interval (CI) of the average ratings, plotted based on the regression coefficients.

The scale score 5 is defined as "allowed to practice EPA only under reactive/
Validity Arguments for EPA-Based Assessment

Academic Medicine, Vol. 96, No. 7S / July 2021 Supplement

on-demand supervision with supervisor immediately available, all findings double checked" (Table 2; score = 5; entrustment 3a). Using the desired threshold for entrustment to allow a student to move into the transition phase to residency (cutoff ≥ 5), Figures 1 and 2 demonstrate that the average time to meet this threshold varies by EPA. Many students met or exceeded the cutoff for entrustment (≥ 5) at approximately 3 months (Figures 1 and 2) for EPA 5 (Document a clinical encounter in the patient record) and EPA 6 (Provide an oral presentation of a clinical encounter), at 4 months for EPA 1 (Gather a history and perform a physical examination) and EPA 4 (Enter and discuss orders and prescriptions), and 5 months for EPA 2 (Prioritize a differential diagnosis following a clinical encounter) and EPA 3 (Recommend and interpret common diagnostic and screening tests). By 7 months, all students had met or exceeded the cutoff score (≥ 5).

By 9 months, many students had even met or exceeded the cutoff score (≥ 6) for entrustment corresponding to “allowed to practice EPA only under reactive/on-demand supervision available and key findings double checked.” The 95% CI for the entrustment cutoff score of 6 on each EPA shows that some students had not yet achieved this performance level by the end of the LIC, with the percentage varying by EPA (Figures 1 and 2: EPA 1 = 5%, EPA 2 = 40%, EPA 3 = 0%, EPA 4 = 0%, EPA 5 = 0%, and EPA 6 = 8%).

**Discussion**

In the present study, we employed rating data on EPAs gathered from undergraduate medical students to model their learning curves. We had posited that students’ learning curves would follow Thurstone’s theory of negative exponential growth and that doing so would provide evidence of construct validity. We found that both the raw data plotted over time and the fitted curves did follow a negative exponential function. The negative exponential growth following baseline performance demonstrated a latent phase to a maximal slope, and subsequently decelerating growth, in accordance with theory. Moreover, all students met or exceeded the entrustment cutoff score of 5 (“allowed to practice EPA only under reactive/on-demand supervision available and all findings double checked”). Additionally, there was considerable variability on achieving various levels of entrustment across 6 EPAs. Taken together, these results provide validity evidence of the Core EPA assessment program implemented in the EPAC program at the University of Minnesota. Such an investigation is important for both theoretical and practical reasons. Theoretically, the present study confirmed that some EPAs are more difficult to demonstrate performance at a predetermined level of competence than others. The average time to attain the required level of competence for graduation varied between 3 and 5 months depending on the EPA. The origins of these differences probably include student factors, EPA factors, and curricular factors.

From the student perspective, the developmental trajectory suggested by the RIME model (Reporter, Interpreter, Manager, Educator) is consistent with our findings. EPAs that would match the reporter function (oral presentation and clinical documentation) were the first that students reached threshold performance on, while those requiring interpretation (Prioritize a differential diagnosis) and/or management (Enter and discuss orders and prescriptions and Recommend and interpret common diagnostic and screening tests) required longer time frames to reach the same threshold. EPA factors that might influence time to reach threshold levels of performance include complexity and ease of observation. In the former case, interpreting diagnostic and screening tests is a more complex task than an oral presentation or...
documentation of a clinical encounter. In the case of ease of observation, tasks such as oral presentations are part of the workflow of almost every team, while observing a student performing a history and/or physical exam requires preplanning. Finally, some of the findings may be curricular in nature. Students may enter the clerkship phase more practiced and assessed in some EPAs, such as oral presentations, than others, such as entering orders.

The progression of student performance on the EPAs provides important information for purposes of course design, instruction, and assessment. Practically, since it appears to be more difficult for students to meet or exceed the cutoff score for entrustment for EPA 2 (Prioritize a differential diagnosis following a clinical encounter) and EPA 3 (Recommend and interpret common diagnostic and screening tests), it makes sense to ensure that there is sufficient time, curriculum/educational experiences, assessment, faculty development, and feedback in these areas to allow for competency to be achieved.

**Limitations**

EPAC is a pilot program employing a small number of students and we were only able to provide analyses of 6 EPAs that had sufficient numbers of observations for regression analyses. By making the unit of analyses for the learning curves the number of observations rather than the individual students, however, we were able to conduct multivariate regression analyses to develop theoretically fitted curves.

**Conclusions**

The analyses of the more than 2,400 assessments in the EPAC program's first 4 years for the Core EPAs 1–6 are consistent with long-standing learning theory and provide construct validity evidence for the EPAC program of assessment.

**Funding/Support:** None reported.

**Other disclosures:** None reported.

**Ethical approval:** This study was determined to be exempt from review by the University of Minnesota–Twin Cities Institutional Review Board in September 2018.

---

**References**


Appendix 1

Hierarchical Regression Analyses of Education in Pediatrics Across the Continuum (EPAC) Data

Even though there are some missing values in EPAs 1–6, there were enough data to conduct an analysis of multivariate hierarchical linear modeling (MHLM). The analysis was conducted via HLM 7. In these analyses, for each time point (months) in an EPA, indicator variables were used, and for the available data points, these variables were coded as dummy variables, 1 and 0.

As the main purpose of these analyses was to examine the growth of the students throughout the LIC period, a polynomial regression (i.e., several independent variables) was hypothesized as a representative of the relationship between time-variable (months) and students’ average ratings. Therefore, statistical significance of the quadratic term in the polynomial regression equations at 95% confidence interval level was examined.

The main purpose of the present analysis was to examine if the growth curve has a linear or curvilinear structure parallel to learning curve theory. Thus, curvilinear relationships were examined at the group level. However, since there are potential individual differences between students apart from the impact of the program, hierarchical linear modeling (HLM) was applied. Even though HLM approaches can model both the fixed and random effects, due to small sample size with some missing values at several time points, only fixed effects of the MONTH variable were estimated. Hierarchical models as an example is given for any EPA:

Level-1 Model:

\[ EPA_1 = \pi_0 + \pi_1 \times (MONTH) + \pi_2 \times (MONTH^2) + \epsilon \]

Level-2 Model:

\[ \pi_0 = \beta_{00} + r_{0i} \]
\[ \pi_1 = \beta_{10} \]
\[ \pi_2 = \beta_{20} \]

Accordingly, subscripts \( i \) and \( t \) represent individuals and time, respectively:

- \( \pi_0 \) is the group mean of the ratings for an EPA when the assessments were conducted for the first time in June
- \( \pi_1 \) is the linear term of the growth trajectory parameter that shows how much rating was increased on average from June to July
- \( \pi_2 \) is the quadratic term that indicates how much the rating increased when each month changed.

Since each of the monthly ratings was nested within the students, we have a 2-level regression model. In such cases, level-1 growth parameters may have both fixed and random effects. Fixed effects are the impact of an independent variable on a dependent variable on average (at the group level). Random effects represent how variable this effect is across the individuals. Since in this study we have a small sample size, random effects of the month variables were not used. However, since ignoring the natural differences between individuals may obscure the significance of the regression coefficients, random effect \( (r_{ij}) \) was only included to the intercept term \( (\pi_0) \).